EXHIBIT A194

Journal of Modern Applied Statistical Methods
Fall 2002, Vol. 1, No 2, 473-478

1538 – 9472/02/$30.00

On The Misuse Of Confidence Intervals For Two Means In Testing For The Significance Of The Difference Between The Means

George W. Ryan Steven D. Leadbetter
Centers For Disease Control And Prevention

Comparing individual confidence intervals of two population means is an incorrect procedure for determining the statistical significance of the difference between the means. We show conditions where confidence intervals for the means from two independent samples overlap and the difference between the means is in fact significant.

Key words: Overlapping confidence intervals, significance tests, statistical tests of significance, tests for differences of means

Introduction

When conducting a hypothesis test on the difference between two means (i.e., H_o: $\mu_1 - \mu_2 = 0$) or the special case of the difference between two proportions (i.e., H_o: $p_1 - p_2 = 0$) from two independent samples, some practitioners, researchers, and students may be tempted to compare the confidence intervals for the two individual means to determine the statistical significance of the difference. If the individual confidence intervals overlap, one might conclude, in error, that the means do not differ because of this overlap.

George W. Ryan is a Mathematical Statistician, Office of Statistics & Programming (OSP), National Center for Injury Prevention and Control (NCIPC), CDC, Atlanta, Georgia (e-mail: gyr0@cdc.gov). He is a graduate of Texas A&M University with over 20 years experience as a statistician in the federal government. Steven D. Leadbetter is a Mathematical Statistician, OSP, NCIPC, CDC, Atlanta, GA (e-mail: SLeadbetter@cdc.gov). He received an M.S. in Applied Statistics from North Dakota State University and has more than 18 years experience as a statistician with the federal government. The authors thank Marcie-jo Kresnow and Scott R. Kegler for helpful comments.

We say that confidence intervals for means μ_1 and μ_2 computed from sample means $\bar{x}_1$ and $\bar{x}_2$, where $\bar{x}_1 \leq \bar{x}_2$, overlap if the upper bound on $\bar{x}_1$ exceeds the lower bound on $\bar{x}_2$. This misinterpretation of confidence intervals occurs widely in practice (Schenker & Gentleman, 2001); many researchers and even some statisticians mistakenly believe it. Accordingly, we consider the separate confidence intervals associated with the individual hypothesis tests for μ_1 and μ_2 (i.e., $H_{0_1}: \mu_1 = 0$ and $H_{0_2}: \mu_2 = 0$) and the implications of attempting to test the hypothesis H_o: $\mu_1 - \mu_2 = 0$ in terms of the individual confidence intervals associated with H_{0_1} and H_{0_2}.

Examples of overlapping confidence intervals for means that differ significantly are provided by Nelson (1989) and Barr (1969). Assuming a common known population variance, Nelson (1989) and Barr (1969) show that when given sample means from two normally distributed populations, the appropriate confidence interval for testing the hypothesis H_o: $\mu_1 - \mu_2 = 0$ is based on the difference of the sample means, $\bar{x}_1 - \bar{x}_2$. We generalize this result to include the assumption of unequal sample variances and the special case of two proportions.

Methodology

Statistically Significant Difference of Two Means

Consider the case of independent random samples of size n_1 and n_2 from two populations with sample

means $\bar{x}_1$ and $\bar{x}_2$ and variances s_1^2, s_2^2. For simplicity, assume the population variances are equal and the populations are either normally distributed or the samples are sufficiently large so the assumptions of the Student's *t*-test are satisfied for the hypothesis tests and confidence intervals (Woodward, 1999). (This assumption will avoid any unnecessary complications with the distribution of the test statistic when the population variances are unequal.) The two sample means differ significantly at the .05 alpha level if the difference $|\bar{x}_1 - \bar{x}_2|$ exceeds about 2 standard errors of the difference of the means (i.e., $|\bar{x}_1 - \bar{x}_2| \geq 2s_{\bar{x}_1 - \bar{x}_2}$).

For simplicity and clarity, because this discussion is in an instructional context, we use the quantity 2 as a sufficiently close approximation to the critical value of the Student's *t*-distribution at the .05 alpha level, which for large sample sizes will be close to the standard normal distribution critical value of 1.96. How can this difference hold if the individual confidence intervals for μ_1 and μ_2 overlap? If the confidence intervals overlap and the sample means $\bar{x}_1$ and $\bar{x}_2$ differ significantly, then (from Figure 1 below), it is necessary that $s_{\bar{x}_1} + s_{\bar{x}_2} > s_{\bar{x}_1 - \bar{x}_2}$. That is, the sum of the individual standard errors must exceed the standard error of the difference of the means.

An estimate of $\sigma^2_{\bar{x}_1 - \bar{x}_2}$ is given by $s^2_{\bar{x}_1 - \bar{x}_2} = s^2(1/n_1 + 1/n_2)$, where $s^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$ is an estimate of σ^2 obtained by pooling s_1^2 and s_2^2 (Woodward, 1999). To be significant at the .05 alpha level, the difference in means $|\bar{x}_1 - \bar{x}_2|$ must equal or exceed

$$2s\sqrt{1/n_1 + 1/n_2} \qquad (1)$$

But for the confidence intervals to overlap, the difference between the means must be less than

$$2(s_1/\sqrt{n_1} + s_2/\sqrt{n_2}) \qquad (2)$$

Accordingly, if $|\bar{x}_1 - \bar{x}_2|$ is greater than or equal to (1) but less than (2), the difference of the means is significant and the individual confidence intervals overlap.

Example. The following data for two independent samples is taken from Woodward (1999). For the first sample, $n_1 = 39$, $\bar{x}_1 = 6.168$, and $s_1 = 0.709$; for the second sample, $n_2 = 11$, $\bar{x}_2 = 6.708$, and $s_2 = 0.803$. The computed *t*-statistic for the test of the hypothesis H_o: $\mu_1 - \mu_2 = 0$ is $t(48) = -2.17$ (Woodward, 1999, p. 78) with a resulting p-value of .0351, indicating significance at the .05 alpha level. The 95% confidence intervals for μ_1 and μ_2 are (5.938, 6.398) and (6.169, 7.247), respectively. Accordingly, the sample means $\bar{x}_1$ and $\bar{x}_2$ differ significantly ($p = .0351$) yet the confidence intervals overlap. Moreover, note the conditions from (1) and (2) above and in Figure 1 are satisfied; i.e., $2s_{\bar{x}_1} + 2s_{\bar{x}_2} > |\bar{x}_1 - \bar{x}_2| \geq 2s_{\bar{x}_1 - \bar{x}_2}$; for this example, $.711 > .540 > .498$.

Statistically Significant Difference of Two Proportions

Two independent proportions, p_1 and p_2, may also be used to illustrate that overlapping confidence intervals do not imply nonsignificance of the observed difference. We now assume the samples are sufficiently large so that p_1 and p_2 (and hence their difference) are normally distributed. To be significant at the .05 alpha level, the difference $|p_1 - p_2|$ in the proportions must equal or exceed

$$2\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2} \qquad (3)$$

However, individual confidence intervals for p_1 and p_2 will overlap if $|p_1 - p_2|$ is less than

$$2(\sqrt{p_1(1-p_1)/n_1} + \sqrt{p_2(1-p_2)/n_2}) \qquad (4)$$

using the quantity 2 as a sufficiently close approximation to the appropriate value (1.96) of the standard normal distribution. For $0 < p_1, p_2 < 1$, and $n_1, n_2 > 1$, the quantity (3) will always be less than (4). So, it could happen that $|p_1 - p_2|$ is greater than or equal to (3) but less than (4), in which case the difference between the proportions would be significant and the confidence intervals would overlap.

Figure 1. Necessary conditions for overlapping 95% confidence intervals for two sample means differing significantly (using the quantity 2 as a sufficiently close approximation to the appropriate critical values of the Student's *t*-distribution).

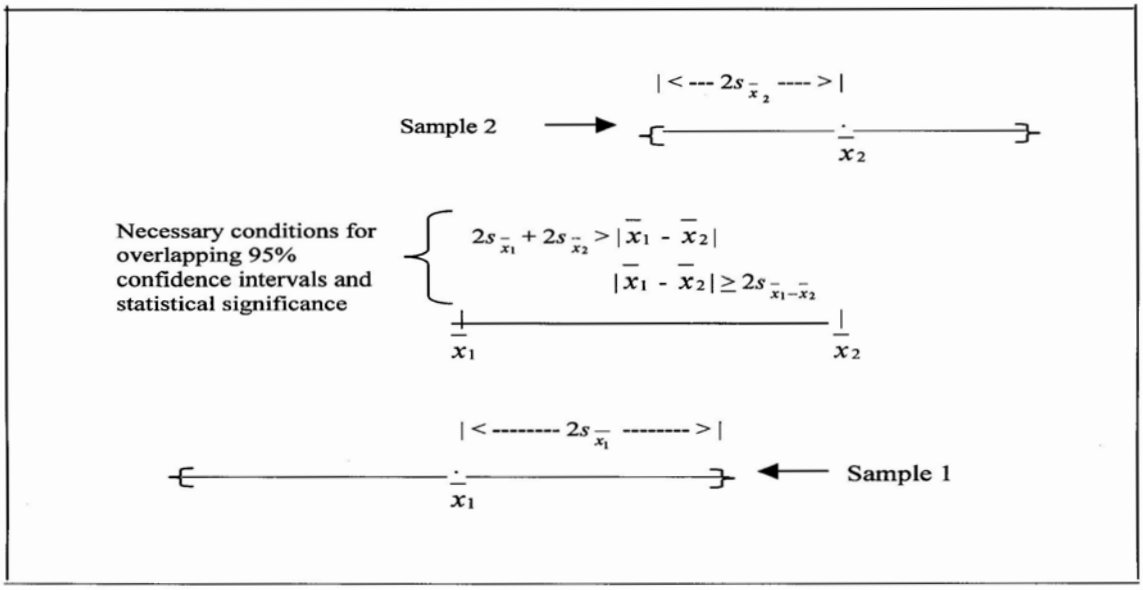
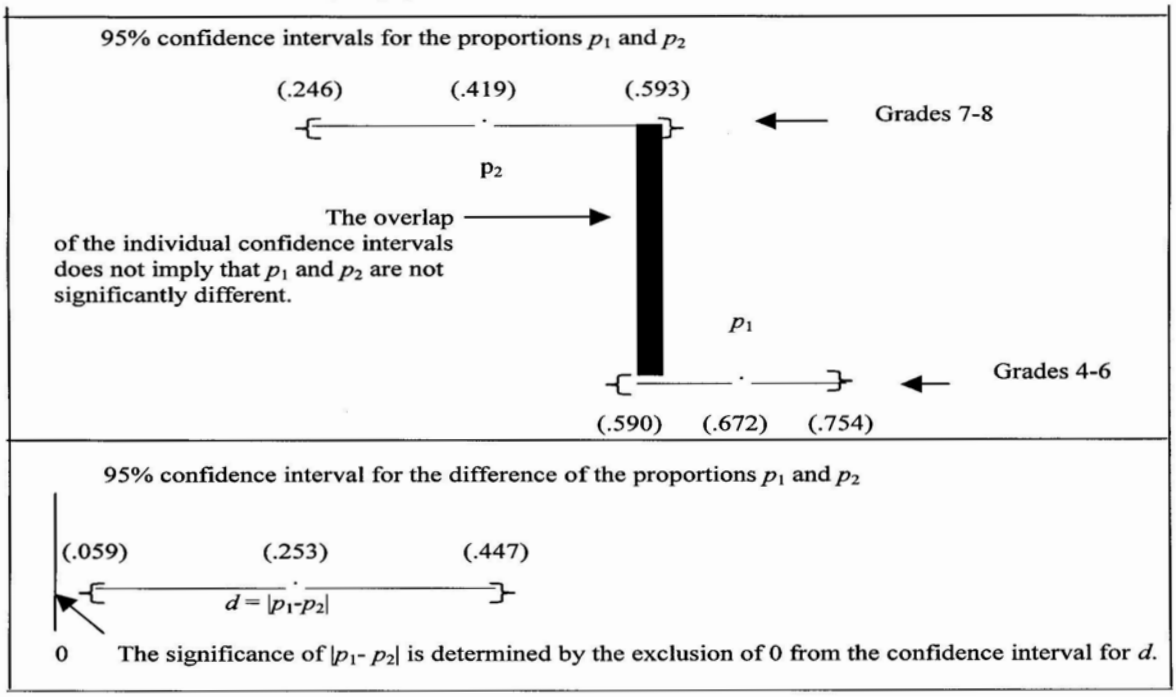


Figure 2. Texas Bicycle Helmet Study Data. Example: 95% confidence intervals for proportions of students agreeing (p_1 in grades 4-6, p_2 in grades 7-8) that "helmets should be worn" and the 95% confidence interval for $d = |p_1 - p_2|$.



Results

The Texas Bicycle Helmet Study (Logan, Leadbetter, & Gibson, 1998) provides an example of two independent proportions p_1 and p_2 with overlapping confidence intervals and a significant difference between the proportions. Elementary and middle school students were surveyed over three time periods to assess their attitudes on such issues as helmet use, school rules, and social acceptability of bicycle helmets. In this example, let p_1 be the proportion of students in grades 4 - 6 in survey period 3 who agree that students "must wear helmets" and p_2 the corresponding proportion of students in grades 7 - 8 (see Figure 2 above). We are interested in testing H_o: $p_1 = p_2$. What result is obtained by observing the individual 95% confidence intervals? How does this result compare with the hypothesis test?

The upper bound of the confidence interval for p_2 (.593) is greater than the lower bound for p_1 (.590), leading some to conclude incorrectly that the observed difference $p_1 - p_2$ is not significant. However, dividing the difference of the proportions (.253) by the standard error of the difference (.098) results in a test statistic of $z = 2.58$, which corresponds to a significance probability (p-value) of .0099. As shown previously, the individual confidence intervals overlap even though p_1 and p_2 differ significantly at the .05 alpha level provided $|p_1 - p_2|$ is less than twice the sum of the individual standard errors of p_1 and p_2. In this example, p_1 and p_2 differ significantly, but the individual confidence intervals overlap as the difference $p_1 - p_2$ (.253) is less than twice the sum of the individual standard errors (2(.042 + .089) = .262).

Of course, the proper interpretation of hypothesis testing in the context of confidence intervals consists (using the present example) of the estimated difference $d = p_1 - p_2$ with *its* associated lower and upper bounds to see if *that* confidence interval includes zero (see Figure 2) (Woodward, 1999). For any significance level, failure of the associated confidence interval to "cover" zero will always indicate significance in the corresponding hypothesis test. To correctly interpret the relationship between confidence intervals and hypothesis tests, one needs to use the confidence interval of the difference.

Conclusion

Our purpose has been to show that an overlap of individual confidence intervals for two means or proportions does not necessarily indicate that the difference between the means is nonsignificant. The proper interpretation of confidence intervals is important because of their increased use in recent years as an inferential tool in preference to traditional hypothesis testing (Chow, 1996). In disciplines such as medicine (Gardner & Altman, 1986), epidemiology (Savitz, Tolo, & Poole, 1994), education (Nix & Barnette, 1998), and psychology (Krantz, 1999), many believe that confidence intervals are more meaningful and easier to interpret than tests of significance.

This erroneous use of individual confidence intervals to determine the significance of the difference between two means could lead one to fail to reject the hypothesis of no difference when the difference is indeed significant. This misuse of individual confidence intervals results in an overly conservative test (Schenker & Gentleman, 2001). In the Texas Bicycle Helmet Study, which used .05 as the stated alpha level, the actual significance probability (p-value) was .0099, indicating a significant difference of means.

The erroneous interpretation of overlapping confidence intervals would lead one to conclude otherwise. The potential for misinterpretation is even more profound if the observations are taken from a sample of paired data since the standard error of the difference (between the observations in each pair) can be considerably smaller (assuming the sample means are positively correlated) than the standard errors of the means from the individual samples (Woodward, 1999). Using the individual confidence intervals here to test the hypothesis H_o: $d = 0$ (d being the difference within each paired observation) would be an exceedingly conservative procedure.

To indicate how individual 95% confidence intervals can overlap even when the means differ significantly, we generated confidence intervals for two proportions p_1 and p_2 for a range of sample sizes. Using values of $p_1 = .65$ and $p_2 = .40$ (chosen because they are comparable to the values in the previous example) and, for simplicity, equal size samples from each

population (i.e., $n_1 = n_2 = n$), we computed confidence intervals for p_1 and p_2. Percent overlap is defined as the ratio of the amount of overlap of the confidence intervals to the difference $p_1 - p_2$. For sample sizes ranging from 30 to 57 from each population, the individual confidence intervals overlap and the two proportions differ significantly (see Figure 3).

For $n < 30$, the individual confidence intervals overlap, but the difference of the proportions is no longer significant at the .05 alpha level. For $n > 57$, the proportions are significantly different, but the confidence intervals no longer overlap. It is within the range of sample sizes from 30 to 57 (for the selected values of p_1 and p_2) that one could erroneously conclude that the difference $p_1 - p_2$ is significant on the basis of overlapping confidence intervals. As the percent overlap decreases, so too does the significance probability (see Figure 3). Accordingly, the consequences of misinterpretation are greater as the overlap becomes smaller. In the example in Figure 2, the percent overlap is (.593 - .590) / (.672 - .419), or 1.2%, but the significance probability, as previously noted, is .0099.

Note that for any value n selected within the range (30, 57) in Figure 3 (next page) for equal sample sizes ($n_1 = n_2 = n$), the difference $p_1 - p_2$ (.25) will be greater than expression (3) and less than (4), the conditions previously noted for overlapping 95% confidence intervals for two significantly different proportions.

Why does this problem persist? Some users may be accustomed to viewing graphical and other displays of data, such as results of multiple range tests, in which overlapping segments of output do indicate nonsignificant differences. They may jump to the erroneous conclusion that overlapping confidence intervals imply that the difference of the means is nonsignificant. Another notion that may contribute to the belief that overlapping confidence intervals imply a nonsignificant difference is the case of nonoverlapping confidence intervals for proportions from two independent samples (Centers for Disease Control and Prevention, 1995).

In the case of two proportions, from the conditions noted in (3) and (4), the sum of the individual standard errors always exceeds the standard error of the difference. It then follows that if the confidence intervals do not overlap, the difference of the proportions is indeed significant. This fact may lead some to conclude that two proportions do not differ significantly if their confidence intervals do overlap.

So what do the individual confidence intervals say about the difference between the means? These intervals are statements only about the variability of each individual estimate; they say nothing about their difference. To determine the significance of the difference in the context of a confidence interval, lower and upper bounds for the difference can be computed quite routinely once the standard error of the difference between the means has been obtained. Only by looking at the lower and upper confidence limits for this difference (see Figure 2) and noting whether the interval includes (or excludes) zero, can one determine the statistical significance of the difference.

References

Barr, D. R. (1969). Using confidence intervals to test hypotheses. *Journal of Quality Technology, 1,* 256-258.

Centers for Disease Control and Prevention. (1995). *Healthy people 2000 statistical notes.* Atlanta, GA.

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility.* Thousand Oaks, CA: Sage Publications.

Gardner, M. J. & Altman, D. G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *Statistics in Medicine, 292,* 746-750.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 44,* 1372-1381.

Logan, P., Leadbetter, S., & Gibson, R. E. (1998). Evaluation of a bicycle helmet giveaway program – Texas, 1995. *Pediatrics, 101,* 578-582.

Nelson, L. S. (1989). Evaluating overlapping confidence intervals. *Journal of Quality Technology, 21,* 140-141.

Nix, T. W. & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools, 5,* 3-14.

Savitz, D. A., Tolo, K-A., & Poole, C. (1994). Statistical significance testing in the American Journal of Epidemiology. *American Journal of Epidemiology, 139,* 1047-1052.

Schenker, N. & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55,* 182-186.

Woodward, M. (1999). *Epidemiology: Study design and data analysis.* Boca Raton, FL: Chapman & Hall/CRC.

Figure 3. Percent overlap of confidence intervals for p_1 and p_2 and significance probabilities ($30 \leq n \leq 57, p_1 = .65, p_2 = .40$).